

# Scatterplot with Line Tutorial: Driving Safety (NYT)

All files in this tutorial are available at:

[http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/)

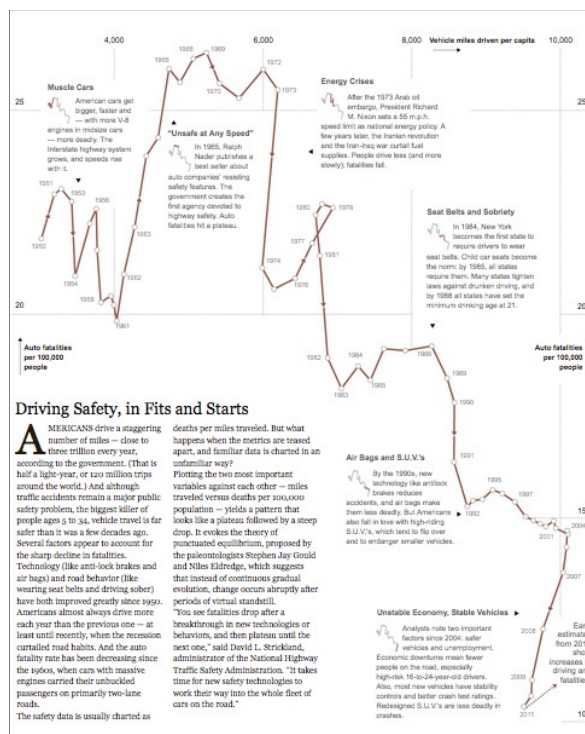
This step-by-step tutorial shows how to recreate a graph from the New York Times, published on Sept 17, 2012, titled *Driving Safety, in Fits and Starts* — you can see the original story here:

<http://www.nytimes.com/interactive/2012/09/17/science/driving-safety-in-fits-and-starts.html>

The graph plots miles driven per capita and traffic fatalities per 100,000 people. Each data point is also the year being plotted, and the cool thing is that the line sometimes snakes backwards (when the miles driven are fewer year-to-year). Even though that makes it look like going backwards in time, each year is really not quantitative data, it's rather qualitative; think of the years as any other category (not values), like the states in the US for example. See also this other data gleaning tutorial, again from the NYT, that uses a similar time-tracking technique:

[http://unixlab.sfsu.edu/~trogu/523/2016/tutorials/data\\_gleaning\\_tutorial\\_d3\\_json\\_csv.pdf](http://unixlab.sfsu.edu/~trogu/523/2016/tutorials/data_gleaning_tutorial_d3_json_csv.pdf)

The original driving safety graph looks like this:



## (1) Find the data

On Wikipedia I found a data set about traffic fatalities in the US going back to the beginning of the 20th Century: [https://en.wikipedia.org/wiki/Motor\\_vehicle\\_fatality\\_rate\\_in\\_U.S.\\_by\\_year](https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year)

It was just a table in the web page, so I simply selected all the rows and columns and pasted the selection into Excel. You have to trial-and-error as sometimes this trick does not work well. Also delete any rows and text that are not data. Keep only the header row with the column headings, and all the needed rows (I started at 1921 and ended at 2017, although the NYT graph only started at 1950).

## (2) Clean the data set and save as a .CSV file

You can do a lot of cleaning up in Excel, such as getting rid of the comma for the thousands separator.

See the original file here:

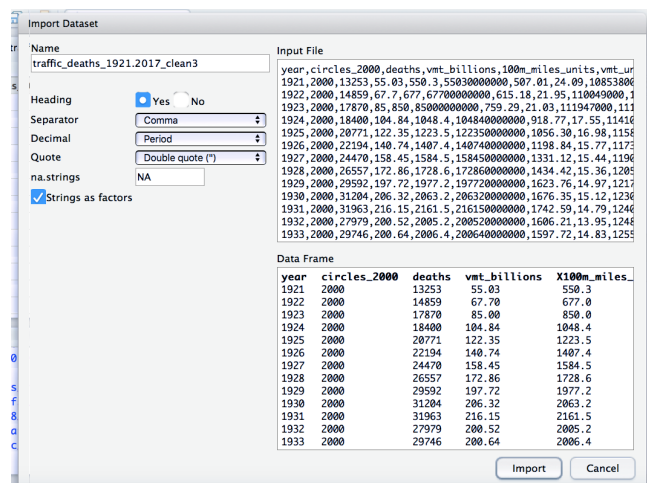
[http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/traffic\\_deaths\\_1900-2017.xlsx](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/traffic_deaths_1900-2017.xlsx)

I simplified the column header names, adding underscores for word spaces (R will put a dot otherwise to replace the spaces.) I added a few columns to get absolute number of deaths in units and fatality rate per 100K. Although I did not need it, I kept the percentage change column, got rid of the “%” symbol, and duplicated the column using only “general” numbers. This makes the number a fraction in decimals; then added another column where I multiplied by 100 to get a simple percentage number again.

This is the new Excel file: [http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/traffic\\_deaths\\_1921-2017\\_clean.xlsx](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/traffic_deaths_1921-2017_clean.xlsx)

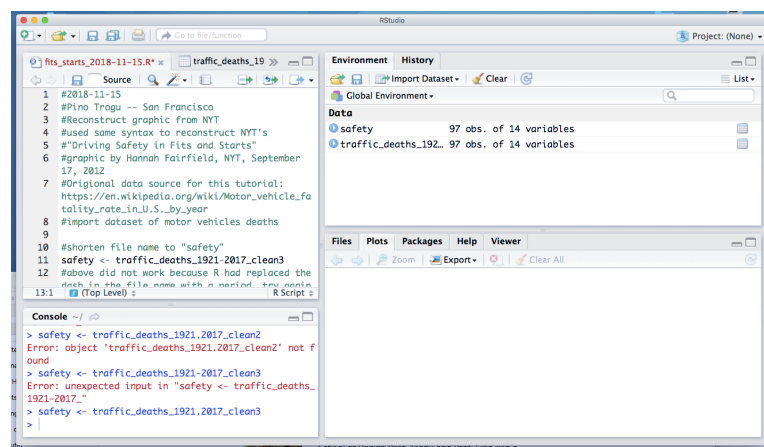
Save the file as CSV but then open it in TextWrangler, NotePad, or other text editor to make sure that the data is clean. With the text editor I also changed the tab separators to comma separators. Show invisibles to do this. This is the text only file (CSV — comma separated value) after cleaning it up with a text editor.  
[http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/traffic\\_deaths\\_1921-2017\\_clean3.csv](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/traffic_deaths_1921-2017_clean3.csv)

Use this data set file for the tutorial. The file is called: [traffic\\_deaths\\_1921-2017\\_clean3.csv](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/traffic_deaths_1921-2017_clean3.csv)



### (3) Import the data set into RStudio

Import the dataset (Environment —> Import dataset —> From text file). Select Heading: Yes (note that R appends an “X” to any header name that begins with a number.) Also, if the file name has any special characters (in this case a dash between the two years), R will substitute a dot in its place (see upper left “Name” in the screenshot.) Note also that in the data set I created a “fake data” column called “circles\_2000”. I will need this later to generate circles all of the same size.

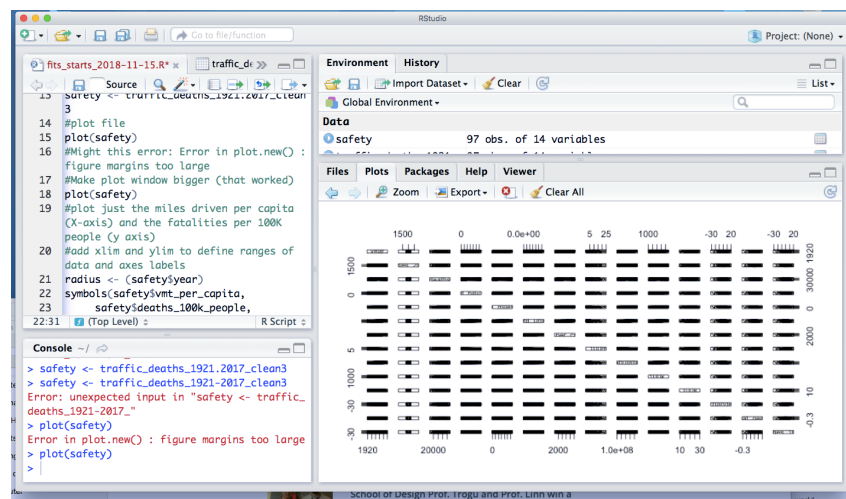


### (4) Shorten the file name to

“safety”. Download and open or cut-paste the R file into RStudio. The file is here:

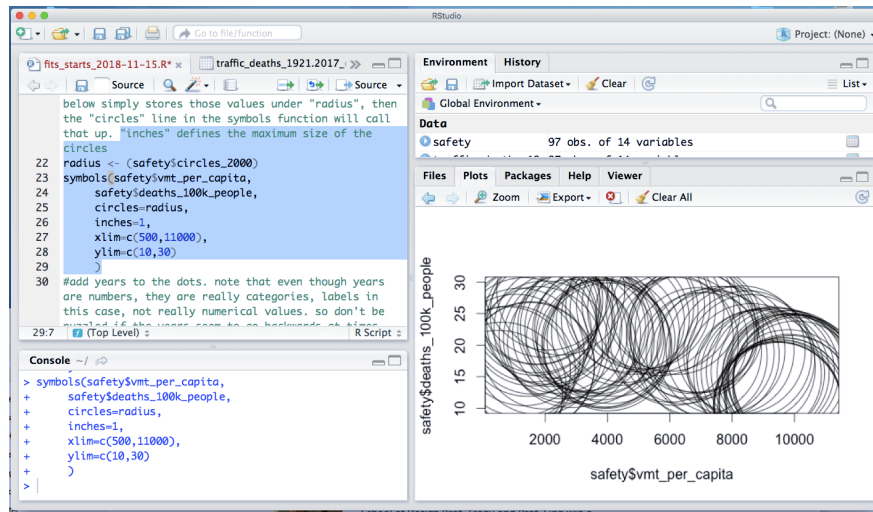
[http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/fits\\_starts\\_2018-11-15.R](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/fits_starts_2018-11-15.R)

Shorten the name but note the error because the file name has the dash instead of the dot. Use the dot in the name to fix it.



### (5) Plot “safety”.

This gives a lot of different combinations (pairs for X and Y using all possible combinations of columns in the data set.)



**(6) Plot VMT (vehicle miles traveled) per capita and deaths per 100K people.**

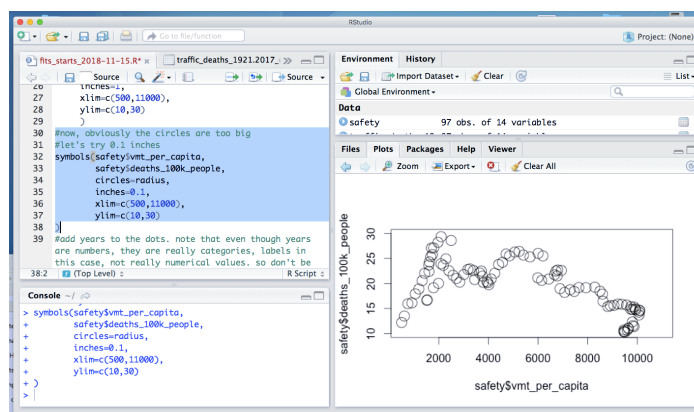
Since we want to replicate the NYT graphic, we want to render the data points just for those two variables (miles on X-axis and deaths on Y-axis).

```
radius <- (safety$circles_2000)
symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=1,
        xlim=c(500,11000),
        ylim=c(10,30)
)
```

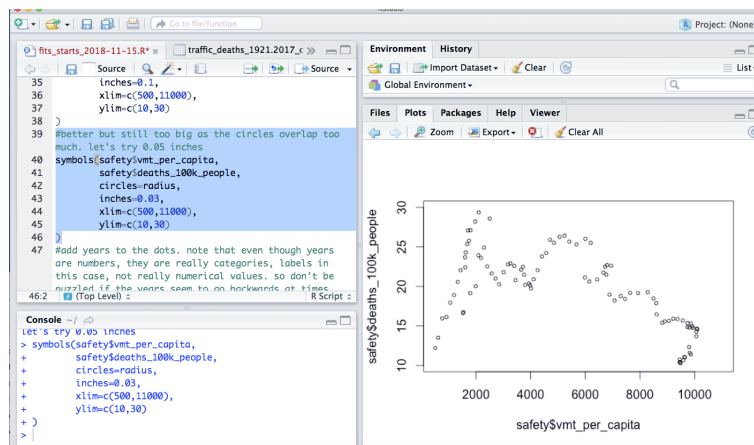
Let's look at the code:

`radius` is shorthand for the data in the `circles_2000` column. We want to generate circles and this will give us circles of the same size (since every row is the same in the fake data column: 2000), located on the X and Y given by the two other variables. The `symbols` function specifies where the circles will be placed (`vmt_per_capita` for X and `deaths_100k_people` for Y). `xlim` and `ylim` define the ranges and the labels for the axes. Look at the dataset for the min and max for the two variables and make the range a little wider on both end, using rounded off numbers. Run `summary` in R to quickly see the min and max for each column.

One inch is obviously too big. Try 0.1 inch. But even with that, circles are still a little too big as there is too much overlap between them, so in the next try we will make them 0.03 inches



```
symbols(safety$vmt_per_capita,
        safety
$deaths_100k_people,
        circles=radius,
        inches=0.1
xlim=c(500,11000),
ylim=c(10,30)
)
```

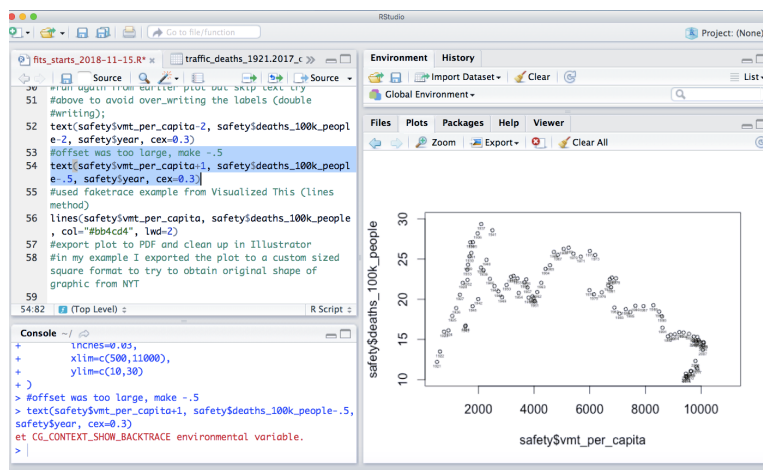


```

symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=0.03,
        xlim=c(500,11000),
        ylim=c(10,30)
)

```

Now add the names (years) to each dot. Try different sizes using `cex`. We can make the text bigger later in Illustrator. Note the offset values `+1` and `-.5`: the names are not centered on dots.

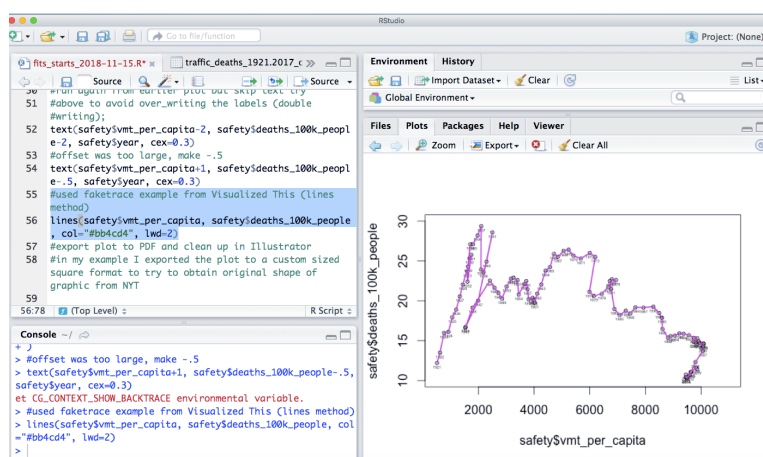


```

text(safety$vmt_per_capita+1,
     safety$deaths_100k_people-.5,
     safety$year,
     cex=0.3
)

```

Now draw a line connecting all the circles (years). Don't be surprised if the line sometimes snakes backwards, that's when people drove less than the previous year.



```

lines(safety$vmt_per_capita,
      safety$deaths_100k_people,
      col="#bb4cd4",
      lwd=3
)

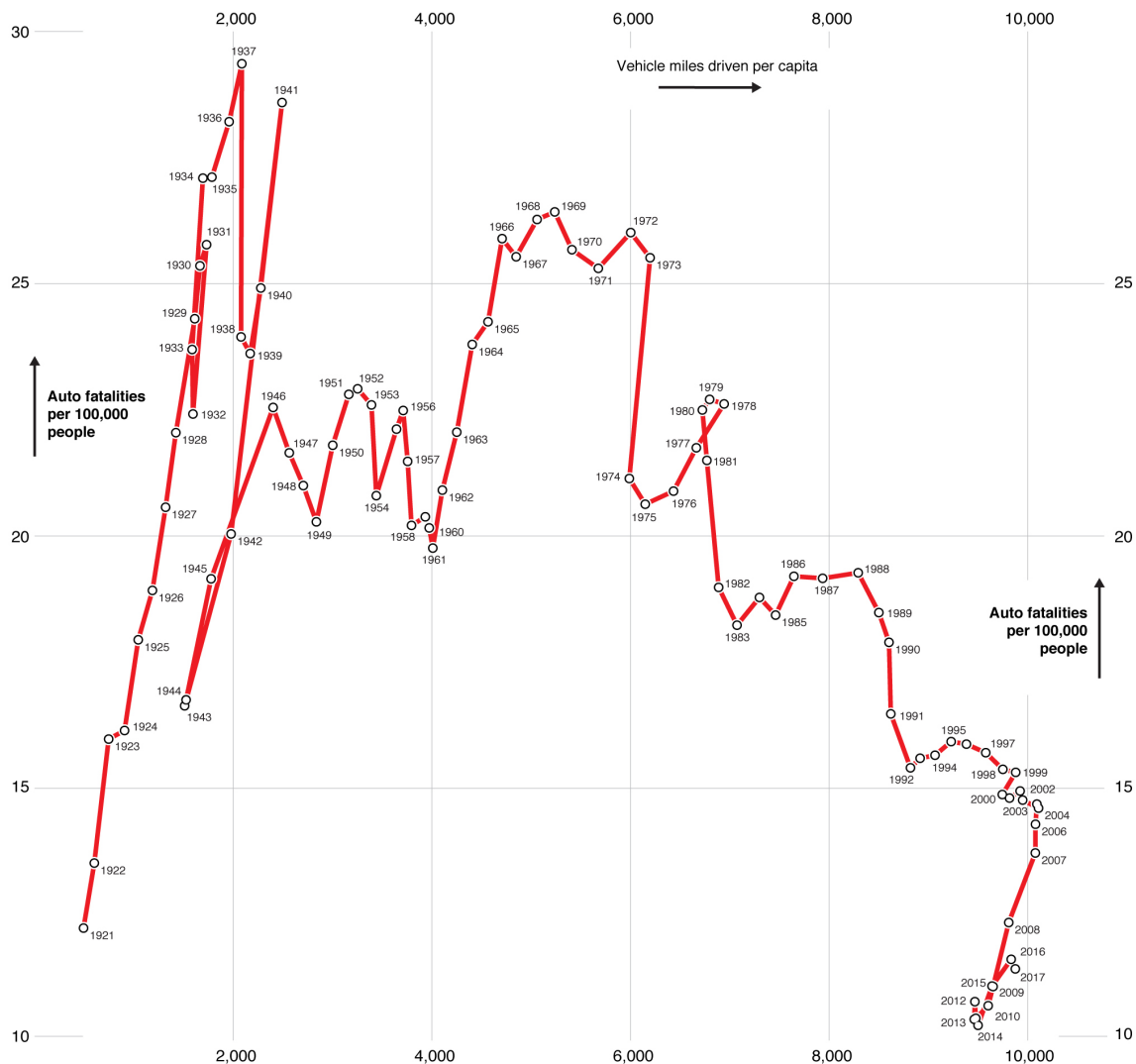
```

Note that the graph is very close to the original in the NYT. The original started in 1950 so our new line shows about thirty extra years on the left, from 1921 to 1950. Note how steep the line is in the beginning, with large

increases in the percentage of fatalities. Export the plot to PDF specifying a square format and open the file in Illustrator. Note: the code in the R file includes a couple of other plots with different variables, tested

just for fun. But from here on we focus on cleaning up the PDF using Illustrator and trying to match the original graph from NYT. From the original I did not include the annotations, the arrow points within the line, or the thumbnail version of each line segment, focusing on each major change noted by a small caption near each small images. All these things are very important, so make sure to include such information (the annotation layer) when doing your own graph.

**(1)** Open file in Illustrator. **(2)** Select all and release clipping mask (Object —> Clipping mask —> Release) to be able to select and delete unwanted elements like the outer box etc. **(3)** Try to group different elements: all text as a group; all circles, etc. This can be done using the Select Same tool after clicking on the desired element. **(4)** Unfortunately we cannot enlarge the circles all at once, as a group, because the distance between them would also increase. So I did a few hacks with many circles on top of each other: a middle layer circle with a very thick black border; a top layer with a smaller white fill and border, so a thin black border overall is the result; finally a bottom layer with an even thicker white border which separates the circles from the zig-zag line under all the dots. Make the main line red and of the desired thickness. See the final illustration below: the overall effect is for the line and circles to look like a road on a map, going from town to town. You can get the Illustrator file here: [http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/driving\\_safety\\_1921\\_2017\\_2.ai](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/driving_safety_1921_2017_2.ai)





This is the full R code used in this tutorial. File name: [fits\\_starts\\_2018-11-15.R](#)

It can be downloaded here:

[http://online.sfsu.edu/trogu/523/2018/tutorials/nyt\\_driving\\_safety/fits\\_starts\\_2018-11-15.R](http://online.sfsu.edu/trogu/523/2018/tutorials/nyt_driving_safety/fits_starts_2018-11-15.R)

```
#2018-11-15
#Pino Trogu -- San Francisco
#Reconstruct graphic from NYT
#used same syntax to reconstruct NYT's
#"Driving Safety in Fits and Starts"
#graphic by Hannah Fairfield, NYT, September 17, 2012
#Original data source for this tutorial:
#https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year
#import dataset of motor vehicles deaths
#in Environment --> Import dataset --> From text file
#shorten file name to "safety"
safety <- traffic_deaths_1921-2017_clean3
#above did not work because R had replaced the dash in the file name with a
period, try again:
safety <- traffic_deaths_1921.2017_clean3
#plot file
plot(safety)
#Might get this error: Error in plot.new() : figure margins too large
#Make plot window bigger (that worked)
plot(safety)
#plot just the miles driven per capita (x-axis) and the fatalities per 100k
people (y axis)
#add xlim and ylim to define ranges of data and axes labels
#use circles method but define the radius of the circles first. as this is taken
from the data, use the "circles_2000" column which has the same value (2000) for
each row. This is just a hack to be able to draw circles and size them as needed.
So the first line below simply stores those values under "radius", then the
"circles" line in the symbols function will call that up. "inches" defines the
maximum size of the circles
radius <- (safety$circles_2000)
symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=1,
        xlim=c(500,11000),
        ylim=c(10,30)
        )
#now, obviously the circles are too big
#let's try 0.1 inches
symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=0.1,
        xlim=c(500,11000),
        ylim=c(10,30)
        )
```

```

#better but still too big as the circles overlap too much. let's try 0.05 inches
symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=0.05,
        xlim=c(500,11000),
        ylim=c(10,30)
)
#Try still smaller: inches=0.03
symbols(safety$vmt_per_capita,
        safety$deaths_100k_people,
        circles=radius,
        inches=0.03,
        xlim=c(500,11000),
        ylim=c(10,30)
)
#add years to the dots. note that even though years are numbers, they are really
categories, labels in this case, not really numerical values. so don't be puzzled
if the years seem to go backwards at times
text(safety$vmt_per_capita, safety$deaths_100k_people, safety$year, cex=0.2)
#make year label a little bigger and offset from dots
#run again from earlier plot but skip text try
#above to avoid over_writing the labels (double #writing);
text(safety$vmt_per_capita-2, safety$deaths_100k_people-2, safety$year, cex=0.3)
#offset was too large, make -.5
text(safety$vmt_per_capita+1, safety$deaths_100k_people-.5, safety$year, cex=0.3)
#used faketrace example from visualized This (lines method)
lines(safety$vmt_per_capita, safety$deaths_100k_people, col="#bb4cd4", lwd=3)
#export plot to PDF and clean up in Illustrator
#in my example I exported the plot to a custom sized square format to try to
obtain original shape of graphic from NYT

*****

#for fun, now plot total miles vs total population
#change xlim and ylim to reflect new ranges
#look at the columns in the dataset to find these values or run "summary"
summary(safety)
#it will give min and max values for each column (displays in console window in
RStudio); however if the values are too big they will be shortened. in that case,
go back to the dataset and simply look up the correct values.
plot(safety$vmt_units,
     safety$population,
     xlim=c(55000000000,3500000000000),
     ylim=c(110000000,330000000)
)
#well, that was not very interesting...
#for more fun, now plot population vs fatalities percentage change

```

```
#change xlim and ylim to reflect new ranges
plot(safety$population,
     safety$times_100,
     xlim=c(110000000,330000000),
     ylim=c(-30,20)
)
#add connecting lines
lines(safety$population,
     safety$times_100,
     xlim=c(110000000,330000000),
     ylim=c(-30,20)
)
#add year labels
text(safety$population,
     safety$times_100-1,
     safety$year,
     cex=0.3
)
#that plot shows the death percentage changes for every year, but it's not as
revealing as the original plot as the NYT did it.
```



